ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Relating compound toxicity to molecular structure using machine learning

Master Thesis

Robin Bosshard

October 16, 2023

Advisors: Dr. Eliza Harris, Dr. K. Arturi, Lilian Gasser
Department of Computer Science, ETH Zürich

Abstract

Abstract goes here.

Contents

Chapter 1

Introduction

intro

1.1 The Challenge of Environmental Pollution

Over the past few decades, the surge in environmental pollution by chemical compounds has been driven by industrial processes, agricultural practices, mobility sector, households, and various other factors, leading to significant ecological and health concerns. While these chemicals can enhance our high living standards and comfort of modern society, they can also pose risks and negatively affect chronicaly or acutely both our health and the environment. Toxic substances threaten wildlife but also makes our air, soil, drinking water and food supply less safe. The EU currently maintains comprehensive chemical regulations, however, it is anticipated that global chemicals production will double by 2030. Moreover, the widespread utilization of chemicals, including their inclusion in consumer goods, is expected to expand further. Even though there are over 275 million known chemical compounds registered by the Chemical Abstracts Service (CAS), merely a tiny fraction of them undergo close scrutiny using conventional methods and even less is known about their toxicity profiles and adverse health effects on our organsims.

Building upon the European Green Deal, the 8th Environment Action Programme, guiding European environmental policy until 2030, reinforces the EU's goal of sustainable living within planetary limits, with a vision extending to 2050. One of its key 2030 objectives is a zero-pollution commitment, covering air, water, and soil, prioritizing the well-being of EU citizens. Notably, the European Commission published a sustainability-focused chemicals strategy [1], aligning with the EU's zero-pollution ambition with one of the objectives to minimize concerning substances by either substituting or phasing them out wherever feasible. Consequently, the urgent need to monitor and effectively assess the hazards associated with the daily entering of thousands

of poorly understood chemicals into our environment becomes increasingly evident.

1.2 The Imperative for Prioritization and Toxicity Assessment

Modern analytical techniques, notably nontarget high-resolution mass spectrometry (NTS HRMS/MS), have undeniably transformed our capacity to detect and quantify environmental pollutants. HRMS/MS can detect a variety of human-made pollutants and environmental contaminants within samples taken from the environment, often with uncertain toxicity profiles. These compounds are assessed based on factors such as abundance and fragmentation data (MS1 and MS2). However, the endeavor to identify compounds and characterize their toxicity remains a resource-intensive and time-consuming process. This challenge is further compounded by the scarcity of reference standards, hindering comprehensive elucidation. Traditionally, the prioritization of unidentified compounds rely on signal intensity as a guiding metric. Unfortunately, this approach falls short in delivering an accurate assessment of environmental exposures, as it tends to overlook the crucial toxicological dimension. Consequently, substances with the potential for severe ecological consequences, such as endocrine-disrupting compounds, frequently evade detection due to their low abundance, despite their high toxicity. Therefore, there is an urgent need for prioritization strategies that incorporate the toxicity and ecological impact more effectively.

1.3 The Promise of Machine Learning in Toxicity Prediction

In the past few years, machine learning has emerged as a transformative force in the field of toxicology, particularly in the realm of high-throughput toxicity prediction. High-throughput screening (HTS) has revolutionized the way we assess toxicity by allowing thousands of in vitro bioassays to be conducted rapidly. This high-throughput approach, coupled with advancements in robotics and automated analysis, has generated vast volumes of toxicity data, paving the way for more comprehensive assessments of chemical compounds. Together with the advent of machine learning, this has enabled the development of predictive models that can accurately predict the toxicity of compounds based on their chemical structure. These models can be trained on large datasets of compounds with known toxicity profiles, allowing them to learn the underlying patterns and relationships between chemical structure and toxicity. Once trained, these models can be used to predict the toxicity of new compounds, even if they have not been tested in the lab. This

approach has the potential to significantly reduce the time and cost of toxicity assessment and plays an important role in the prioritization of compounds for further testing.

1.4 MLinvitroTox: A Novel Approach

In response to the pressing need for a more efficient and comprehensive assessment of environmental contaminants, MLinvitroTox [1], an innovative machine learning framework was introduced. MLinvitroTox leverages molecular fingerprints extracted from fragmentation spectra (MS2), signifying a fundamental shift in how we forecast the toxicity of the myriad unidentified HRMS/MS features. The framework leverages streamlined machine learning techniques to predict the compounds bioactivity in numerous toxicity endpoints. The toxicity database encompasses nearly 300 target-specific and 90 cytotoxic endpoints sourced from ToxCast/Tox21 data.

1.5 Objectives and Significance

The primary objective of this research is to enhance the prediction of compound toxicity, particularly in aquatic environments. Our aim is to provide a more accurate and streamlined method for identifying potential environmental hazards, ultimately contributing to the preservation of aquatic ecosystems. By employing customized molecular fingerprints and robust models, we have achieved significant advancements in the prediction of toxic endpoints, making it possible to foresee potential harm with sensitivities exceeding 0.95. Notably, our use of SIRIUS molecular fingerprints and xboost (Extreme Gradient Boosting) models, complemented by the Synthetic Minority Oversampling Technique (SMOTE) for data imbalance, has yielded consistently successful results. Furthermore, we have validated the effectiveness of MLinvitroTox by applying it to MassBank spectra, demonstrating an average balanced accuracy of 0.75 in predicting toxicity.

1.6 Thesis Structure

This thesis is structured to provide a comprehensive understanding of the development, validation, and application of MLinvitroTox. In the following chapters, we will delve into the technical intricacies, showcase the framework's efficacy through validation on real-world data, and present the practical implications of our research in mapping toxicologically relevant pollution in aquatic environments.

1.7 The Relevance of the Special Issue

Our research is part of the special issue titled "Data Science for Advancing Environmental Science, Engineering, and Technology." This special issue emphasizes the critical role of data science in addressing environmental challenges, aligning perfectly with our mission to enhance environmental monitoring through innovative machine learning techniques.

1.8 A Glimpse into the Future

In the subsequent sections, we will explore the intricacies of our groundbreaking MLinvitroTox framework, aiming to provide a solution to the complex issue of identifying and assessing environmental contaminants swiftly and accurately. Our journey begins with a comprehensive overview of the methodology and development process, setting the stage for a deeper exploration of its capabilities and implications in subsequent chapters.

Let us now embark on this exciting journey into the world of MLinvitroTox and its potential to revolutionize our understanding of environmental toxicity, while harnessing the power of high-throughput toxicity prediction through machine learning.

Chapter 2

Literature Review

2.1 Background

2.2 Context

Chapter 3

Material and Methods

3.1 Invitrodb

The most recent release of the ToxCast's (Toxicity Forecaster) database, referred to as invitroDBv3.5, serves as a source of an extensive collection of high-throughput screening (HTS) targeted bioactivity data. This database encompasses information on a total of 9541 compounds, selectively screened across 2205 assay endpoints. This resource originated from the collaboration of two prominent institutions: the United States Environmental Protection Agency (EPA) through its ToxCast program and the National Institutes of Health (NIH) via the Tox21 initiative. Incorporating data collected from diverse research laboratories, this relational database is openly accessible to the public and can be downloaded directly from the official ToxCast website.

3.1.1 Data Overview

Presence Matrix

Consider a collection of m assay endpoints, denoted by $A = \{a_1, a_2, \ldots, a_m\}$ and a set of n compounds represented as $C = \{c_1, c_2, \ldots, c_n\}$. To facilitate data comprehension, we introduce a *presence matrix* $P \in \{0,1\}^{m \times n}$. Rows, indexed by i, represent assay endpoints a_i, while columns, indexed by j, denote presence (1) or absence (0) of compound c_j in those endpoints. Matrix P is sparse due to the selective testing of compounds across different assay endpoints. A compound is considered present in an assay endpoint if it has undergone testing and a corresponding concentration-response series is available. See Figure 3.1 for a visual of the *presence matrix* P covering all assay endpoints and compounds in *invitroDBv3.5*.

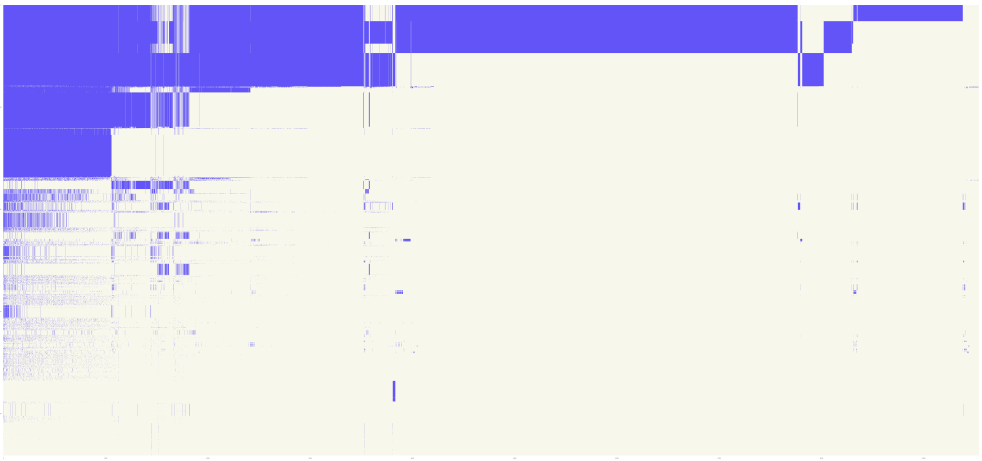

Figure 3.1: The *presence matrix* P covering all assay endpoints and compounds available in *invitroDBv3.5* with $m = 2205$ assay endpoints and $n = 9541$ compounds. The presence matrix is organized by sorting it based on the number of compounds present in each assay endpoint and the compounds are arranged in descending order of their presence frequency. The total count, where $P_{ij} = 1$, indicates the availability of 3 342 377 concentration-response series for downstream analysis.

Subsetting data

We exclusively consider assay endpoints that have been tested with a minimum of 2000 compounds. This criterion ensures the availability of sufficient data for the training of a machine learning model. Refer to Figure 3.2 for a visual representation of the *presence matrix* P, which now encompasses only the resulting subset of all assay endpoints within *invitroDBv3.5*. From now on, we will call this specific subset the data that we will be focusing on for this thesis.

Concentration-Response Series

A *concentration-response series* is represented as a set of k concentration-response pairs:

$$S = \{(conc_1, resp_1), (conc_2, resp_2), \ldots, (conc_k, resp_k)\}$$

For each entry in the presence matrix P with $P_{ij} = 1$, we collect the corresponding concentration-response series S_{ij} for the compound c_j in the assay endpoint a_i. We analyse in total $\sum_{i,j} P_{ij} = 1\,372\,225$ concentration-response series, comprising a sum of $\sum_{i,j} |S_{ij}| = 48\,861\,036$ concentration-response pairs across all compounds and assay endpoints. We get the concentration-response pairs by combining tables mc0, mc1, and mc3 from invitroDBv3.5. We also gather necessary sample information such as well type, row, and

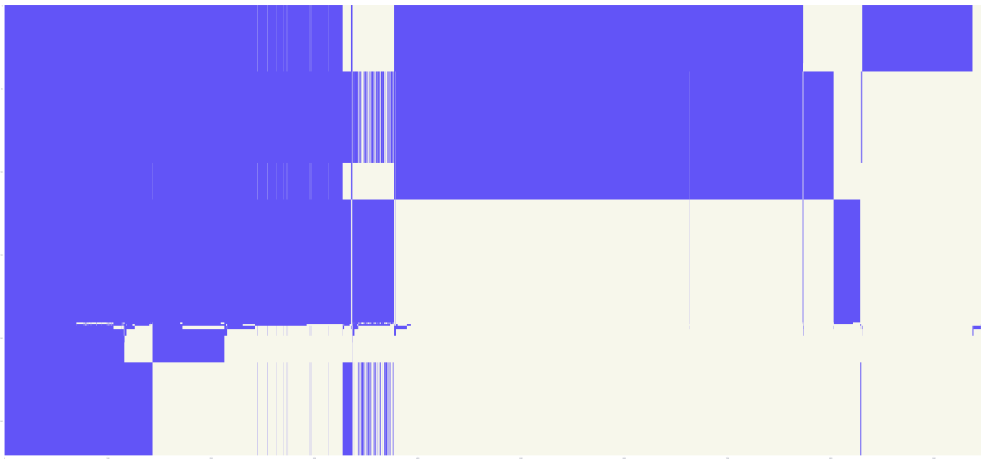

Figure 3.2: The *presence matrix* P covering only the subset of all of assay endpoints available in *invitroDBv3.5*, considered for this thesis, encompassing $m = 271$ assay endpoints and $n = 9456$ compounds. The total count, where $P_{ij} = 1$, indicates the availability of 1 372 225 concentration-response series for downstream analysis.

column index from the assay well-plate. The concentrations are transformed to the logarithmic scale using the unit μM (micromolar), while the responses are normalized to either fold-induction or percent-of-control units. Figure 3.3 showcases a single concentration-response series for some compound tested within an assay endpoint.

Figure 3.3: A concentration-response series for the compound *Estropipate* (DTXSID3023005) in the assay endpoint *TOX21_ERa_LUC_VM7_Agonist* (aeid=788). The series has a total of $k = 45$ concentration-response pairs and is composed of $n_{conc} = 15$ concentration groups, each with $n_{rep} = 3$ replicates.

In this section, we demonstrate the significance of variations in concentration-

response pairs among different compounds and assay endpoints. In practice, concentrations are often subjected to multiple testing iterations, resulting in the formation of distinct concentration groups. Within each concentration group, the number of replicates is indicated by n_{rep}. We introduce the following quantities corresponding to a concentration-response series for a compound c_i in a given assay endpoint a_i:

- $n_{\text{datapoints}_{i,j}}$: the total number of concentration-response pairs ($|S|$)
- $n_{\text{groups}_{i,j}}$: the number of distinct concentrations tested
- $n_{\text{replicates}_{i,j}}$: the number of replicates for each concentration group
- $min_{\text{conc}_{i,j}}$: the lowest concentration tested
- $max_{\text{conc}_{i,j}}$: the highest concentration tested

For an overview of these quantities across the entire set of considered concentration-response series, please refer to Figure 3.4. This figure illustrates the above metrics aggregated by their means, grouped by assay endpoints and compounds.

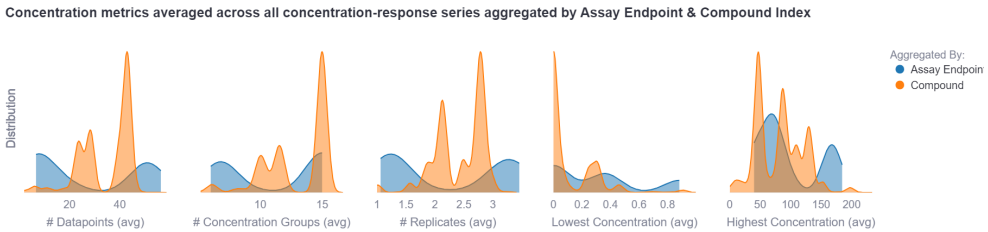

Figure 3.4: Concentration metrics averaged across all concentration-response series aggregated by assay endpoint (blue) and compound (orange). E.g., the first chart shows the distribution on the average number of datapoints across all assay enpoint $a_i \in A$ with $\frac{1}{|C|}\sum_j n_{\text{datapoints}_{i,j}}$ and across all compounds $c_j \in C$ with $\frac{1}{|A|}\sum_i n_{\text{datapoints}_{i,j}}$. The same is done for the other metrics: $n_{\text{groups}_{i,j}}$, $n_{\text{replicates}_{i,j}}$, $min_{\text{conc}_{i,j}}$, and $max_{\text{conc}_{i,j}}$.

3.2 Pytcpl

We introduce pytcpl, a streamlined Python package inspired by the R package tcpl, designed for processing high-throughput screening data. The package primarily focuses on providing essential features such as concentration-response curve fitting and allows for continuous hit-calling for compound bioactivity across diverse assay endpoints, akin to tcplfit2. Invitrodb version 3.5 release can optionally serve as backend database if desired. The package optimizes data storage and provides compressed raw data and metadata from *invitroDB* in Parquet files. This efficient strategy reduces storage needs,

resulting in just 4 GB within the repository—compared to the original 80 GB database. This obviates the need for a cumbersome, large-scale database installation, rendering downstream analysis more accessible and efficient. Our package is crafted to accomodate cusomizable processing steps and facilitate interactive data visualization with curve surfer. Moreover, it empowers Python-oriented researchers to seamlessly engage in data analysis and exploration.

3.2.1 Pipeline

1. Data collection
2. Cutoff determination and filtering (Meet conditions for curve fitting)
3. Curve fitting
4. Hit calling

Data Collection

First, all datapoints are collected from the database and assigned to the concentration response-series belonging to the respective compound in the corresponding assay endpoint.

Curve Fitting

Introduce all candidate fit models, discuss the pros and cons of each model. Discuss the fitting procedure, how the models are fitted, Maximum Likelihood Estimation

Hit Calling

Akaike criterion, probability of being active, etc..

3.2.2 Curve Surfer

Data visualization, overview of what is possible with the tool. Filter by assay endpoint, compound, etc.

3.3 Machine Learning Pipeline

3.3.1 Preprocessing

Subselecting the columns from the output tables generated by pytcpl: DTXSID identifier and continuous hitcall value. The feature inputs to the machine learning model is a molecular structure represented as fingerprint generated from a SMILES string uniquely determined by the compounds DTXSID

identifier. The SMILES string is a linear representation of a compound's molecular structure. The SMILES string is converted to a molecular graph, which is then converted to a feature vector. The feature vector is then used to train a machine learning model. The machine learning model is then used to predict the hitcall value for a given compound. The machine learning pipeline is illustrated in Figure **??**.

3.3.2 Binary Classification

The goal is to predict whether a compound is active or inactive for a given assay endpoint. We can formulate this as a binary classification problem, where the input is the compound's molecular structure fingerprint and the output is the hitcall value binarized by some decision threshold. The hitcall value is rendered to a binary variable, where 1 indicates that the compound is active and 0 indicates that the compound is inactive.

3.3.3 Regression

3.3.4 Massbank Validation

Chapter 4

Results and Discussion

sectionResults sectionEvaluation sectionDiscussion

Chapter 5

Conclusion

We have evidence of a multitude of chemicals being present in the environment and in our bodies and that mixture exposure indeed matters. This knowledge needs to be deepened, and the quantitative contribution of chemicals to compromised health should be better described and translated into regulatory action. As indicated in a scientific opinion paper of the German Federal Environmental Agency (Conrad et al. 2021), the CSS goals may be considered as a moving target. For increasing scientific evidence and improved method for detection and assessment of chemicals, development of new technologies require innovative regulatory, technological and societal reactions. We should be flexible and prepared to take up the scientific challenges and collaborate productively with regulatory institutions to address the identified challenges and modernise chemical risk assessment. This is also in line with the concern of many scientists that chemical pollution and the wide range of adverse effects on human and ecosystem health demand additional efforts on a global scale (Brack et al. 2022; Wang et al. 2021). We see the CSS as a European strategy that, in concert with other initiatives, may open new opportunities to minimise hazardous chemical pollution and thus risks to human health and ecosystems.

Bibliography

[1] K. Arturi and J. Hollender, "Machine learning-based hazard-driven prioritization of features in nontarget screening of environmental high-resolution mass spectrometry data," *Environmental Science & Technology*, vol. 0, no. 0, null, 0, PMID: 37279189. DOI: 10.1021/acs.est.3c00304. eprint: https://doi.org/10.1021/acs.est.3c00304. [Online]. Available: https://doi.org/10.1021/acs.est.3c00304.

Appendix A

Appendix

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.